Machine Vision and Applications (2016) 27:275–286
DOI 10.1007/s00138-016-0748-8

ORIGINAL PAPER

Articulated tracking with manifold regularized particle filter

Adam Gonczarek[1] · **Jakub M. Tomczak**[1]

Received: 23 March 2015 / Revised: 10 December 2015 / Accepted: 25 December 2015 / Published online: 2 February 2016

Abstract In this paper, we investigate articulated human motion tracking from video sequences using Bayesian approach. We derive a generic particle-based filtering procedure with a low-dimensional manifold. The manifold can be treated as a regularizer that enforces a distribution over poses during tracking process to be concentrated around the low-dimensional embedding. We refer to our method as *manifold regularized particle filter*. We present a particular implementation of our method based on back-constrained gaussian process latent variable model and gaussian diffusion. The proposed approach is evaluated using the real-life benchmark dataset *HumanEva*. We show empirically that the presented sampling scheme outperforms sampling-importance resampling and annealed particle filter procedures.

Keywords Articulated motion tracking · Manifold regularization · Generative approach · Back-constrained gaussian process latent variable model

List of symbols

$\mathcal{I}$	Set of all available images from all cameras
$\mathcal{I}_t$	Set of all available images in the tth moment
$\mathcal{I}_{1:T}$	Set of all available images from the first to the Tth moment
$\mathbf{x}$	A human body configuration
$\mathbf{x}_t$	A human body configuration in the tth moment
$\mathbf{x}_{1:T}$	The whole trajectory of body configurations from the first to the Tth moment
$\mathbf{z}$	A human body configuration in the low-dimensional manifold coordinate system
$\mathbf{z}_t$	A human body configuration in the low-dimensional manifold coordinate system in the tth moment
$\mathbf{z}_{1:T}$	The whole trajectory of human body configurations in the low-dimensional manifold coordinate system from the first to the Tth moment
$\delta(\cdot)$	The Dirac delta function
S^{I}	A binary silhouette obtained by subtracting background from the input image I
$\mathrm{S}^{\mathrm{I}}(\mathbf{x})$	A binary silhouette obtained by projecting a body model generated from a given pose $\mathbf{x}$ onto the image I
$\lvert\cdot\rvert$	Matrix determinant
$\mathrm{tr}(\cdot)$	Matrix trace
$\lVert\cdot\rVert_F$	The Frobenius norm
$\mathbf{I}_{D\times D}$	The D-dimensional identity matrix
$k(\cdot,\cdot)$	A kernel function
$\mathbf{K}$	A kernel matrix

1 Introduction

Articulated human motion tracking from video image sequences is one of the most challenging computer vision problems for the past two decades. The basic idea behind this issue is to recover a motion of a complete human body basing on the image evidence from a single or many cameras. Moreover, it is assumed that the motion tracking is performed without any additional devices, e.g., color or electromagnetic markers. The human motion tracking system can be

✉ Adam Gonczarek
adam.gonczarek@pwr.edu.pl

[1] Department of Computer Science, Wrocław University of Technology, wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland

applied in many everyday life areas, see [9,15,18]. Giving some examples, it may be used in control devices for human–computer interaction, surveillance systems detecting unusual behaviors, dancing or martial arts training assistants, support systems for medical diagnosis.

During last years, a lot of effort has been put in solving the human motion tracking issue. However, excluding some minor cases, the problem remains open. There are several reasons worth mentioning that make the issue challenging. First, there is a huge variety of different images corresponding to the same pose that may be obtained. This is caused by diversity of human wear and appearance, changes in lighting conditions, camera noise, etc. Second, an image lacks depth information which makes impossible to obtain three-dimensional pose from two-dimensional images. Moreover, one has to handle different types of occlusions including self-occlusions and occlusions caused by external environment. Finally, efficient exploration of the space of all possible human poses is troublesome because of high-dimensionality of the space and its non-trivial constraints.

To date, however, several conceptually different approaches have been proposed to address the human motion tracking problem. They can be roughly divided into two groups.

In the first group, discriminative methods are used to model directly the probability distribution over poses conditioned on the image evidence. This approach is usually composed of two parts where feature extraction is followed by prediction using multivariate regression model. To obtain informative features simple techniques like binary silhouettes [1,14] as well as more sophisticated descriptors like histogram of oriented gradients or a HMAX model [3] were adapted. As a regression model a whole spectrum of different techniques were used, e.g., ridge regression and support vector machines [1], mixture of experts [10], gaussian processes [3], kernel information embedding [14].

On the contrary, in the second group a generative approach is used to model separately the prior distribution over poses and the likelihood of how well a given pose fits to the current image. Pure generative modeling assumes that one tries to model the true pose space and uses Bayesian inference to combine this prior knowledge with the image evidence to estimate the current pose. Within this group of methods the two important branches have evolved.

In the first one, kinematic-tree models are used to represent the body pose. Since it is straightforward to render a 3D body model using this representation, the likelihood is usually computed by comparing the difference between the given images and body model projections. The main effort in this branch of methods is to design an appropriate pose prior. Many strategies were applied here, from simple limits on joint movements [20] to more advanced models trying to capture the manifolds of human poses embedded in high-dimensional pose space, e.g., gaussian process latent variable model (GPLVM) [25,27], gaussian process dynamical model [26], mixtures of factor analyzers [13], hierarchical hidden Markov model [17], restricted Boltzmann machines [24]. Eventually, we can predict the body pose by finding *maximum a posteriori* (MAP) estimator or use a fully Bayesian approach by computing the complete posterior distribution over poses. The latter approaches usually take advantage of particle filters to approximate the posterior.

In the second branch of methods part-based models are used for pose recovery [2,4,22,23,29]. Here, we assume that all body parts are modeled individually. More flexible priors are used to cover many possible relative positions between parts. The main effort is put in constructing rich likelihood models that are required to detect individual body parts. In general, inference is based on searching MAP estimate using for example dynamic programming [29] or Branch and Bound methods [22,23]; however, there are some individual cases where a fully Bayesian inference is used, see [5]. Finally, part-based models are mainly applied to 2D pose estimation.

In this paper, we present a novel fashion of involving information about low-dimensional embedding in the pose space into the tracking process that leads to a generic filtering procedure. We use the generative approach based on a kinematic-tree model and Bayesian inference. We propose a particle filter-based algorithm for the filtering problem, which we will refer to as *manifold regularized particle filter*. Finally, we present a dynamics model based on gaussian process latent variable model with back constraints.

The contribution of the paper is fourfold. First, a new class of particle filter algorithms is proposed where a low-dimensional information is involved into the inference process. This allows to utilize full Bayesian reasoning. Second, we present a specific instance of the proposed approach that utilizes the gaussian process latent variable model with back constraints combined with gaussian diffusion. Third, the outlined approach is applied to the articulated human-motion tracking. Fourth, we show empirically that the presented sampling scheme outperforms sampling-importance resampling and annealed particle filter procedures on benchmark dataset HumanEva.

The paper is organized as follows. In Sect. 2 the problem of the human motion tracking is outlined and the proposed filtering procedure is presented. In Sect. 3 the manifold regularized particle filter is proposed. The model of dynamics with low-dimensional manifold is presented in Sect. 5. At the end, the empirical study is carried out in Sect. 6 and conclusions are drawn in Sect. 7.

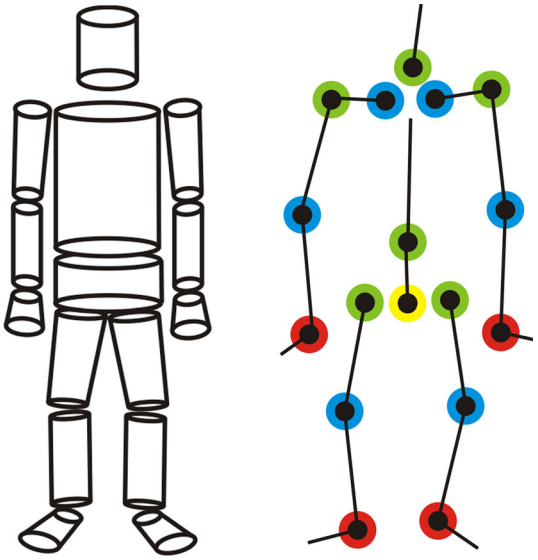


Fig. 1 **a** Human body model represented by articulately connected rigid parts. **b** Kinematic tree representing connections between neighboring rigid parts. *Red*, *blue*, and *green* vertices correspond to joints with one, two and three degrees of freedom, respectively. *Yellow* vertex is the root of the tree

2 Human motion tracking

In this paper, we assume that a human body is represented by a set of articulately connected rigid parts (see Fig. 1a). Each connection between two neighboring elements characterizes a joint and can be described by up to three degrees of freedom, depending on movability of the joint. All connected parts form a kinematic tree with the root typically associated with the pelvis. A common representation of the state of the kth joint uses Euler angles that describe relative rotation between neighboring parts in the kinematic tree (see Fig. 1b). However, we prefer to use quaternions because they can be compared using the Euclidean distance metric. Moreover, if the relative rotations between connected parts in the kinematic tree are small, i.e., in the range between 0 and π then we can reduce quaternion representation to 3-tuple instead of 4-tuple, see [21] or [5] for details.

The set of quaternions for all relative rotations in the kinematic tree together with the global position and orientation in 3D constitutes the minimal set of variables that are used to describe the current pose of the human body, which is denoted by $\mathbf{x}$. It is worth mentioning that $\mathbf{x}$ is usually around 40–50 dimensions, which is one of the fundamental reasons that makes the human motion tracking a difficult problem due to intractability in searching over high-dimensional spaces.

We assume that there are several synchronized cameras that provide video images of a human body from different perspectives. The cameras should be located so that to contribute as much information about the body as possible, i.e., they should register different parts of the scene. We will denote a set of all available images from all cameras by $\mathcal{I}$.

In typical pose estimation problem, we want to estimate the human body configuration $\mathbf{x}$ basing on $\mathcal{I}$. Thus, the key issue is to properly model the conditional distribution $p(\mathbf{x}|\mathcal{I})$. Since it is a multivariate regression model, we can estimate the pose by computing the expected value $\hat{\mathbf{x}} = \mathbb{E}[\mathbf{x}|\mathcal{I}]$. In the generative approach we follow the Bayes rule to inverse the conditional probability:

$$p(\mathbf{x}|\mathcal{I}) \propto p(\mathcal{I}|\mathbf{x})p(\mathbf{x}), \tag{1}$$

and then model the prior $p(\mathbf{x})$ and the likelihood $p(\mathcal{I}|\mathbf{x})$ separately.

We can extend the individual pose estimation problem to tracking of the whole trajectory $\mathbf{x}_{1:T} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ in the pose space. Then, let $\mathcal{I}_{1:T} = \{\mathcal{I}_1, \ldots \mathcal{I}_T\}$ denote the corresponding sequence of available images.

Before giving the formal problem statement, notice that the high-dimensional pose space consists of human body configurations where most of them are unrealistic. Additionally, during specific motions (e.g., walking or running) all degrees of freedom exhibit strong correlations that depend on the current pose. These two remarks yield a corollary that the true trajectories form a low-dimensional manifolds. Therefore, we assume that any pose $\mathbf{x}$ corresponds to a point $\mathbf{z}$ in the coordinate system on the low-dimensional manifold.

Formally, in the generative approach we need to model the joint probability distribution $p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}, \mathcal{I}_{1:T})$. This task is rather difficult unless we assume some conditional independence between variables. Typically, it is assumed that current pose depends only on the current low-dimensional representation, as well as the current observation depends only on the current pose. Temporal dependencies are assumed only between low-dimensional representations [25,26,28]. This is a reasonable assumption if the variance of the conditional distribution $p(\mathbf{x}_t|\mathbf{z}_t)$ is low. However, we have found out that for typical video frame rates (e.g., 60 Hz) the variance of the distribution $p(\mathbf{x}_t|\mathbf{z}_t)$ is usually much higher than the variance of the distribution $p(\mathbf{x}_t|\mathbf{x}_{t-1})$. Hence, this leads to the corollary that human motion formulates continuous trajectories that locally oscillate around the low-dimensional manifold.

Therefore, we indicate that it is more important to model temporal coherence between high-dimensional poses, and low-dimensional representations should be used just to keep the trajectory close to the manifold. The manner how the joint probability distribution is factorized is presented by the probabilistic graphical model in Fig. 2. Notice that the current state $\mathbf{x}_t$ influences future state and future point on the manifold $\mathbf{z}_{t+1}$ which in turn impacts $\mathbf{x}_{t+1}$.

In the Bayesian inference we are interested in calculating the posterior probability distribution for $\mathbf{x}_t$ given images $\mathcal{I}_{1:t}$

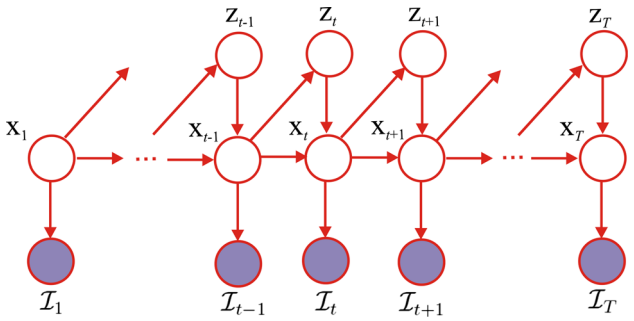


Fig. 2 Probabilistic graphical model presenting how the joint probability distribution $p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}, \mathcal{I}_{1:T})$ factorizes

by marginalizing out all previous poses $\mathbf{x}_{1:t-1}$ and hidden variables $\mathbf{z}_{1:t}$ from $p(\mathbf{x}_{1:t}, \mathbf{z}_{1:t}|\mathcal{I}_{1:t})$ which yields:

$$p(\mathbf{x}_t|\mathcal{I}_{1:t}) = \frac{p(\mathcal{I}_t|\mathbf{x}_t)}{p(\mathcal{I}_t|\mathcal{I}_{1:t-1})} \iint p(\mathbf{z}_t|\mathbf{x}_{t-1}) \times p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}) \mathrm{d}\mathbf{x}_{t-1} \mathrm{d}\mathbf{z}_t, \quad (2)$$

where $p(\mathcal{I}_t|\mathcal{I}_{1:t-1})$ is a normalization constant given by:

$$p(\mathcal{I}_t|\mathcal{I}_{1:t-1}) = \iint p(\mathcal{I}_t|\mathbf{x}_t) p(\mathbf{z}_t|\mathbf{x}_{t-1}) \times p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}) \mathrm{d}\mathbf{x}_{t-1:t} \mathbf{z}_t. \quad (3)$$

We have obtained a filtering procedure that includes information about the low-dimensional manifold and is independent of actual forms of each component. Further, we show that if the filtering procedure is performed in this manner, then choosing relatively simple component models $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$ and $p(\mathbf{z}_t|\mathbf{x}_{t-1})$ leads to very promising results.

3 Manifold regularized particle filter

In the context of the human motion tracking the filtering procedure is intractable usually since we are unable to compute analytically the integral in (2) and the normalization constant (3) except the case where all distributions are gaussian. Hence, an approximation of the posterior should be applied. Typically, sampling methods like particle filter-based techniques are used, in the context of human motion tracking the most popular method is known as *Condensation algorithm* [8]. However, the main disadvantage of this technique is that it requires to generate a huge amount of particles in order to cover a high-dimensional state space. Otherwise, it fails to approximate the true distribution. In order to cover the highly probable areas in the pose space only, an extension called *annealed particle filter* (APF) has been proposed [6]. However, in this method particles tend to be trapped in one or a few dominating local maxima in the posterior distribution. Therefore, the method is non-robust to cases where substantial number of local maxima occurs and thus fails to track the proper trajectory. This usually happens when the image evidence is inconsiderable, e.g., we use noisy likelihood model or small number of cameras.

In this paper, we propose a different approach that modifies the Condensation algorithm by introducing a regularization in a form of the low-dimensional manifold. This filtering procedure operates in the neighborhood of the low-dimensional space where the true poses are concentrated, and thus it guarantees that highly probable regions are covered and the particles are distributed around different local extrema.

In fact the proposed particle-based filtering procedure provides a proxy to the posterior $p(\mathbf{x}_t|\mathcal{I}_{1:t})$ obtained in (2), because we can approximate the distribution as follows:

$$p(\mathbf{x}_t|\mathcal{I}_{1:t}) \approx \sum_{n=1}^{N} \pi(\mathbf{x}_t^{(n)}) \delta(\mathbf{x}_t - \mathbf{x}_t^{(n)}), \quad (4)$$

where $\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(N)} \sim p(\mathbf{x}_t|\mathcal{I}_{1:t-1})$ denote samples from the current prior, $\delta(\cdot)$ is the Dirac delta function, and $\pi(\mathbf{x}_t^{(n)})$ is a normalized form of a single score calculated using the likelihood model $\tilde{\pi}(\mathbf{x}_t^{(n)}) = p(\mathcal{I}_t|\mathbf{x}_t^{(n)})$, so:

$$\pi(\mathbf{x}_t^{(n)}) = \frac{\tilde{\pi}(\mathbf{x}_t^{(n)})}{\sum_{j=1}^{N} \tilde{\pi}(\mathbf{x}_t^{(j)})}. \quad (5)$$

However, in most cases it is troublesome to generate a sample from the prior $p(\mathbf{x}_t|\mathcal{I}_{t-1})$ using standard sampling techniques for directed graphical models since generating $\mathbf{x}_t$ from $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$ is usually intractable. Thus, we introduce an auxiliary distribution $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ from which we can sample effectively. Then, taking advantage of conditional independencies defined by the probabilistic graphical model in Fig. 2, we get:

$$p(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}) = \frac{1}{Z} \tilde{\omega}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t) \times q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}), \quad (6)$$

where $\tilde{\omega}$ are weight coefficients defined as follows:

$$\tilde{\omega}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t) = \frac{p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}, \quad (7)$$

and $q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1})$ is a joint auxiliary distribution:

$$q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t|\mathcal{I}_{1:t-1}) = q(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{z}_t|\mathbf{x}_{t-1}) \times p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1}), \quad (8)$$

and Z is a normalization constant.

Let us assume that we can sample from $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ and $p(\mathbf{z}_t|\mathbf{x}_{t-1})$.[1] Thus, once we have a sample from the previous posterior $p(\mathbf{x}_{t-1}|\mathcal{I}_{1:t-1})$, eventually we can easily generate a sample from the auxiliary joint distribution (8) and further we can use it to approximate the distribution (6). Since integrating out variables in discrete approximations is trivial, we obtain the following expression for the prior:

$$p(\mathbf{x}_t|\mathcal{I}_{1:t-1}) \approx \sum_{n=1}^{N} \omega(\mathbf{x}_t^{(n)}, \mathbf{x}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})\delta(\mathbf{x}_t - \mathbf{x}_t^{(n)}), \tag{9}$$

where coefficients $\omega(\mathbf{x}_t^{(n)}, \mathbf{x}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})$ are obtained by normalizing (7) in analogical manner as in (5).

Eventually, by combining the prior (9) with the image likelihoods we can approximate the posterior distribution (2) using the following formula:

$$p(\mathbf{x}_t|\mathcal{I}_{1:t}) \approx \sum_{n=1}^{N} \xi(\mathbf{x}_t^{(n)}, \mathbf{x}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})\delta(\mathbf{x}_t - \mathbf{x}_t^{(n)}), \tag{10}$$

where the weights are defined as follows:

$$\xi(\mathbf{x}_t^{(n)}, \mathbf{x}_{t-1}^{(n)}, \mathbf{z}_t^{(n)}) = \frac{\tilde{\pi}(\mathbf{x}_t^{(n)})\tilde{\omega}(\mathbf{x}_t^{(n)}, \mathbf{x}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})}{\sum_{j=1}^{N} \tilde{\pi}(\mathbf{x}_t^{(j)})\tilde{\omega}(\mathbf{x}_t^{(j)}, \mathbf{x}_{t-1}^{(j)}, \mathbf{z}_t^{(j)})}. \tag{11}$$

Notice that we introduce the low-dimensional manifold in the manner that the particles are weighted both by the image evidence and coefficients $\tilde{\omega}$.

As a result we obtain a new generic particle-based filtering procedure which takes advantage of the prior given by the low-dimensional manifold. The proposed approach is referred to as *manifold regularized particle filter* (MRPF). The schematic representation of the MRPF is presented in Fig. 3 and the detailed procedure is presented in Algorithm 1. Notice that to mitigate particle degeneration phenomenon, which is a typical problem in particle filter-based algorithms, we apply resampling procedure. For more detailed considerations on the importance of resampling see for example [7]. To avoid confusions we use $\mathcal{X}_t$ and $\overline{\mathcal{X}}_t$ to denote samples before and after resampling, respectively.

4 Likelihood function

The likelihood function $p(\mathcal{I}_t|\mathbf{x}_t)$ aims at evaluating the given human body configuration $\mathbf{x}_t$ corresponds to the set of images

[1] Exemplary models of $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ and $p(\mathbf{z}_t|\mathbf{x}_{t-1})$ will be discussed in Sect. 5.

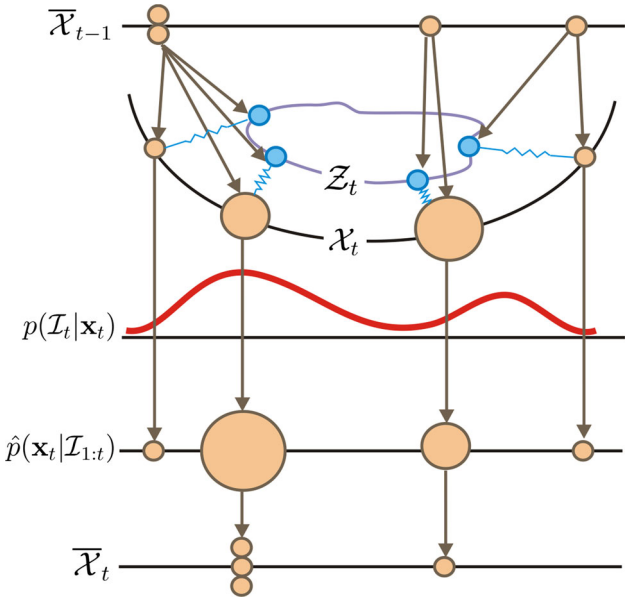


Fig. 3 Schematic illustration of the manifold regularized particle filter

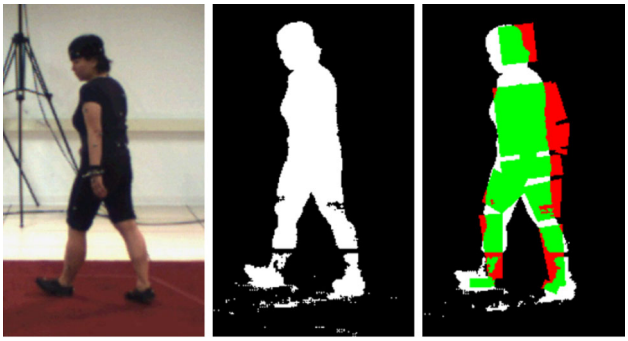
Algorithm 1: Manifold regularized particle filter

Input : initial state $\mathbf{x}_0$, sequence of images $\mathcal{I}_{1:T}$
Output: sequence of state estimates $\hat{\mathbf{x}}_{1:T}$

Duplicate the initial state $\mathbf{x}_0$ and formulate a set: $\overline{\mathcal{X}}_0 = \{\overline{\mathbf{x}}_0^{(1)}, \ldots, \overline{\mathbf{x}}_0^{(N)}\}$;
for $t = 1 : T$ **do**
  Generate a sample $\mathcal{Z}_t = \{\mathbf{z}_t^{(1)}, \ldots, \mathbf{z}_t^{(N)}\}$ using $\mathbf{z}_t^{(n)} \sim p(\mathbf{z}_t|\overline{\mathbf{x}}_{t-1}^{(n)})$;
  Generate a sample $\mathcal{X}_t = \{\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(N)}\}$ using $\mathbf{x}_t^{(n)} \sim q(\mathbf{x}_t|\overline{\mathbf{x}}_{t-1}^{(n)})$;
  Calculate $\tilde{\pi}(\mathbf{x}_t^{(n)})$ using the likelihood model $p(\mathcal{I}_t|\mathbf{x}_t)$;
  Calculate $\tilde{\omega}(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})$ using (7);
  Normalize the weights $\omega(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})$ using (11);
  Calculate the estimate of the state variables $\hat{\mathbf{x}}_t = \sum_{n=1}^{N} \omega(\mathbf{x}_t^{(n)}, \overline{\mathbf{x}}_{t-1}^{(n)}, \mathbf{z}_t^{(n)})\mathbf{x}_t^{(n)}$;
  Resample particles to obtain $\overline{\mathcal{X}}_t = \{\overline{\mathbf{x}}_t^{(1)}, \ldots, \overline{\mathbf{x}}_t^{(N)}\}$ using approximation (10);
end

$\mathcal{I}_t$. We compare images which contain a human body model (see Fig. 1a) projected onto camera views with binary silhouettes obtained from background subtraction procedure, by calculating the difference between them. This model is called *bidirectional silhouette likelihood* [20] and can be defined as follows:

$$-\ln p(\mathcal{I}|\mathbf{x}) = \frac{1}{|\mathcal{I}|}\sum_{\mathrm{I}\in\mathcal{I}}\left\{\frac{1}{|\{\mathrm{S}_{ij}^{\mathrm{I}}(\mathbf{x})=1\}|}\sum_{(i,j)\in\{\mathrm{S}_{ij}^{\mathrm{I}}(\mathbf{x})=1\}}\left(1-\mathrm{S}_{ij}^{\mathrm{I}}\right)\right.$$
$$\left.+\frac{1}{|\{\mathrm{S}_{ij}^{\mathrm{I}}=1\}|}\sum_{(i,j)\in\{\mathrm{S}_{ij}^{\mathrm{I}}=1\}}\left(1-\mathrm{S}_{ij}^{\mathrm{I}}(\mathbf{x})\right)\right\}+\text{const}, \tag{12}$$



(a) **(b)** **(c)**

Fig. 4 **a** An example input image I. **b** Binary silhouette S^{I} obtained in background subtraction procedure. **c** Comparison between S^{I} and binary silhouette $\mathrm{S}^{\mathrm{I}}(\mathbf{x})$ obtained by projecting body model onto image I

where S^{I} denotes binary silhouette obtained by subtracting background from the input image I, and $\mathrm{S}^{\mathrm{I}}(\mathbf{x})$ denotes binary silhouette obtained by projecting a body model generated from a given pose $\mathbf{x}$ onto the image I. Additional constant value in (12) corresponds to the normalizing coefficient in the probability distribution that is independent of the image. We use a simple body model composed of articulately connected cylindrical elements. Analogous model was used in [20]. The idea of calculating the likelihood is presented in Fig. 4.

5 Dynamics model using low-dimensional manifold

We propose to model the dynamics using low-dimensional manifold and a nonlinear dependency. First, we need to learn the low-dimensional manifold. Second, we need to construct a model of dynamics on the low-dimensional manifold, $p(\mathbf{z}_t|\mathbf{x}_{t-1})$, and the model of dynamics in the pose space with the low-dimensional manifold, $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$.

5.1 Learning the low-dimensional manifold

For learning the low-dimensional manifold, we apply the *Gaussian process latent variable model* (GPLVM) [11]. The GPLVM model constitutes a nonlinear dependency between the pose and the low-dimensional manifold as follows:

$$\mathbf{x} = \mathbf{f}(\mathbf{z}) + \boldsymbol{\varepsilon}, \tag{13}$$

where ith function is a realization of the gaussian process [19], $f_i \sim \mathcal{GP}(f|0, k(\mathbf{z}, \mathbf{z}'))$, where k is a kernel function, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\varepsilon}|0, \sigma_z^2\mathbf{I}_{D\times D})$ denotes univariate gaussian noise, where σ_z^2 is variance, and $\mathbf{I}_{D\times D}$ denotes the D-dimensional identity matrix. In this paper, we use the RBF kernel,

$$k(\mathbf{z}, \mathbf{z}') = \beta \exp\left(-\frac{\gamma_z}{2}\|\mathbf{z} - \mathbf{z}'\|^2\right) + \beta_0, \tag{14}$$

where β, β_0, and γ_z are kernel parameters.

We are interested in finding a matrix of low-dimensional variables corresponding to observed poses, i.e., a matrix $\mathbf{Z}$ for observed poses $\mathbf{X}$. Additionally, we want to determine the mapping between the manifold and the high-dimensional space by learning parameters β, β_0 and γ_z, and σ_z^2. The training corresponds to finding the parameters and points on the manifold that maximize the logarithm of the likelihood function in the following form:

$$\begin{aligned}\ln p(\mathbf{X}|\mathbf{Z}) &= \ln \prod_{i=1}^{D} \mathcal{N}(\mathbf{X}_{:,i}|0, \mathbf{K} + \sigma_z^2\mathbf{I}_{T\times T}) \\ &= -\frac{DT}{2}\ln(2\pi) - \frac{D}{2}\ln|\overline{\mathbf{K}}| + \\ &\quad - \frac{1}{2}\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}\mathbf{X}),\end{aligned} \tag{15}$$

where $\mathbf{X}_{:,i}$ denotes ith column of the matrix $\mathbf{X}$, $|\cdot|$ and $\mathrm{tr}(\cdot)$ are matrix determinant and trace, respectively, $\overline{\mathbf{K}} = \mathbf{K} + \sigma_z^2\mathbf{I}_{T\times T}$, and $\mathbf{K} = [k_{nm}]$ is the kernel matrix with elements $k_{nm} = k(\mathbf{z}_n, \mathbf{z}_m)$.

Let us notice that solutions of the maximization $\mathbf{z}_t$ and γ_z can be arbitrarily re-scaled, thus, there are many equivalent solutions. In order to avoid this issue we introduce a regularizer $\frac{1}{2}\|\mathbf{Z}\|_F^2$, where $\|\cdot\|_F$ is the Frobenius norm, and the final objective function takes the form:

$$L(\mathbf{Z}) = \ln p(\mathbf{X}|\mathbf{Z}) - \frac{1}{2}\|\mathbf{Z}\|_F^2. \tag{16}$$

The objective function can be optimized using standard gradient-based optimization algorithms, e.g., scaled conjugate gradient method. Additionally, the objective function is not concave and hence it has multiple local maxima. Therefore, it is important to carefully initialize the numerical algorithm, e.g., by using principal component analysis.

Optimization algorithms need information about the gradient of the objective function. In order to calculate the gradient of (16) w.r.t. $\overline{\mathbf{K}}$ we use the properties of derivatives for matrices [16], which yields:

$$\frac{\partial L(\mathbf{Z})}{\partial \overline{\mathbf{K}}} = -\frac{D}{2}\overline{\mathbf{K}}^{-1} + \frac{1}{2}\overline{\mathbf{K}}^{-1}\mathbf{X}\mathbf{X}^{\mathrm{T}}\overline{\mathbf{K}}^{-1}. \tag{17}$$

Next, the derivative of (16) w.r.t. z_t^i is as follows:

$$\frac{\partial L}{\partial z_t^i} = \mathrm{tr}\left(\left(\frac{\partial L(\mathbf{Z})}{\partial \overline{\mathbf{K}}}\right)^{\mathrm{T}}\frac{\partial \overline{\mathbf{K}}}{\partial z_t^i}\right) - \frac{1}{2}\frac{\partial \|\mathbf{Z}\|_F^2}{\partial z_t^i}, \tag{18}$$

where $\frac{\partial L(\mathbf{Z})}{\partial \overline{\mathbf{K}}}$ is given by (17), and derivatives $\frac{\partial \overline{\mathbf{K}}}{\partial z_t^i}$ are given in the following form:

$$\frac{\partial \bar{k}_z(\mathbf{z}_n, \mathbf{z}_m)}{\partial z_t^i} = \begin{cases} -\gamma_z(k_{nm} - \beta_0)(z_n^i - z_m^i), & t = n \\ \gamma_z(k_{nm} - \beta_0)(z_n^i - z_m^i), & t = m \\ 0, & t \neq n, m \end{cases} \tag{19}$$

and eventually

$$\frac{\partial \|\mathbf{Z}\|_F^2}{\partial z_t^i} = 2z_t^i. \tag{20}$$

We can calculate derivatives w.r.t. β, β_0, γ_z and σ_z^2 analogically.

Notice that the kernel used to determine the covariance function takes high values for points $\mathbf{z}_n$ and $\mathbf{z}_m$ that are close to each other, i.e., they are similar. Moreover, because the points on the manifold are similar, the original poses $\mathbf{x}_n$ and $\mathbf{x}_m$ are similar as well. However, the situation does not hold in the opposite direction. This issue is undesirable in the proposed filtering procedure (2) since the distribution $p(\mathbf{z}_t|\mathbf{x}_{t-1})$ is multi-modal and thus hard to determine. However, this effect can be reduced by introducing *back constraints* that leads to *back-constrained* GPLVM (BC-GPLVM) [12].

The idea behind BC-GPLVM is to define $\mathbf{z}$ as a smooth mapping of $\mathbf{x}$, $\mathbf{z} = \mathbf{g}(\mathbf{x})$. For example, this mapping can be given in the linear form, i.e.,

$$g_i(\mathbf{x}) = \sum_{t=1}^{T} a_{ti} k_x(\mathbf{x}, \mathbf{x}_t) + b_i, \tag{21}$$

where g_i denotes ith component of $\mathbf{z}$, a_{ti}, b_i are parameters, and

$$k_x(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\gamma_x}{2}\|\mathbf{x} - \mathbf{x}'\|^2\right) \tag{22}$$

is the kernel function in the high-dimensional space of poses. We can incorporate the mapping into the objective function, i.e., $z_n^i = g_i(\mathbf{x}_n)$, and then optimize w.r.t. a_{ti} and b_i instead of z_n^i. The application of the back constraints entails closeness of low-dimensional points $\mathbf{z}_t$ if high-dimensional points $\mathbf{x}_t$ are similar.

The big advantage of gaussian processes is tractability of calculating the predictive distribution for new pose $\mathbf{x}_p$ and its low-dimensional representation $\mathbf{z}_p$. The corresponding kernel matrix is as follows:

$$\begin{bmatrix} \overline{\mathbf{K}} & \overline{\mathbf{k}} \\ \overline{\mathbf{k}}^{\mathrm{T}} & \bar{k}_z(\mathbf{z}_p, \mathbf{z}_p) \end{bmatrix}, \tag{23}$$

and finally the predictive distribution [19]:

$$p(\mathbf{x}_p|\mathbf{z}_p, \mathbf{X}, \mathbf{Z}) = \mathcal{N}(\mathbf{x}_p|\boldsymbol{\mu}_p, \sigma_p^2 \mathbf{I}_{D\times D}), \tag{24}$$

where:

$$\boldsymbol{\mu}_p = \mathbf{X}^{\mathrm{T}} \overline{\mathbf{K}}^{-1} \overline{\mathbf{k}}, \tag{25}$$

$$\sigma_p^2 = \bar{k}_z(\mathbf{z}_p, \mathbf{z}_p) - \overline{\mathbf{k}}^{\mathrm{T}} \overline{\mathbf{K}}^{-1} \overline{\mathbf{k}}. \tag{26}$$

5.2 Models of dynamics

Idea of the model $p(\mathbf{z}_t|\mathbf{x}_{t-1})$ is to predict new position on the manifold basing on the previous pose. Therefore, we need a mapping that allows to transform a high-dimensional representation to a low-dimensional one. For this purpose, we apply the back constraints. By adding gaussian noise with the covariance matrix $\mathrm{diag}(\boldsymbol{\sigma}_{x\to z}^2)$ to the back constraints, we obtain the following model of the dynamics on the manifold:

$$p(\mathbf{z}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{z}_t|\mathbf{g}(\mathbf{x}_{t-1}), \quad \mathrm{diag}(\boldsymbol{\sigma}_{x\to z}^2)). \tag{27}$$

On the other hand, the model $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$ determines the probability of the current pose basing on the previous pose and the current point on the low-dimensional manifold. A reasonable assumption is that the model factorizes into two components, namely, one concerning only previous pose, and second—the low-dimensional manifold. This factorization follows from the fact that these two quantities belong to two different spaces and thus are hard to compare quantitatively. Then, the model of the dynamics takes the following form:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) \propto p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_t|\mathbf{z}_t). \tag{28}$$

The first component is expressed as a normal distribution with the diagonal covariance matrix $\mathrm{diag}(\boldsymbol{\sigma}_{x\to x}^2)$:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\mathbf{x}_{t-1}, \quad \mathrm{diag}(\boldsymbol{\sigma}_{x\to x}^2)). \tag{29}$$

The second component is constructed using the mean of the predictive distribution (25) and is disturbed by a gaussian noise with the diagonal covariance matrix $\mathrm{diag}(\boldsymbol{\sigma}_{z\to x}^2)$ which leads to the following model:

$$p(\mathbf{x}_t|\mathbf{z}_t) = \mathcal{N}(\mathbf{x}_t|\mathbf{X}^{\mathrm{T}} \overline{\mathbf{K}}^{-1} \overline{\mathbf{k}}, \mathrm{diag}(\boldsymbol{\sigma}_{z\to x}^2)). \tag{30}$$

It is important to highlight that the training of the parameters $\mathrm{diag}(\boldsymbol{\sigma}_{z\to x}^2)$ has to be performed using a separate validation set which contains data. Otherwise, using the same training set as for determining $\mathbf{Z}$ leads to underestimation of the parameters.

Eventually, let us consider the application of the MRPF (see Algorithm 1) in the context of the outlined models of dynamics. We need to propose the auxiliary distribution $q(\mathbf{x}_t|\mathbf{x}_{t-1})$. In our case it is given in the form (29), i.e., $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathrm{diag}(\boldsymbol{\sigma}_{x\to x}^2))$. Then, the weights $\tilde{\omega}$ are given in the form (30), i.e., $\tilde{\omega}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t) = \mathcal{N}(\mathbf{x}_t|\mathbf{X}^{\mathrm{T}} \overline{\mathbf{K}}^{-1} \overline{\mathbf{k}}, \mathrm{diag}(\boldsymbol{\sigma}_{z\to x}^2))$.

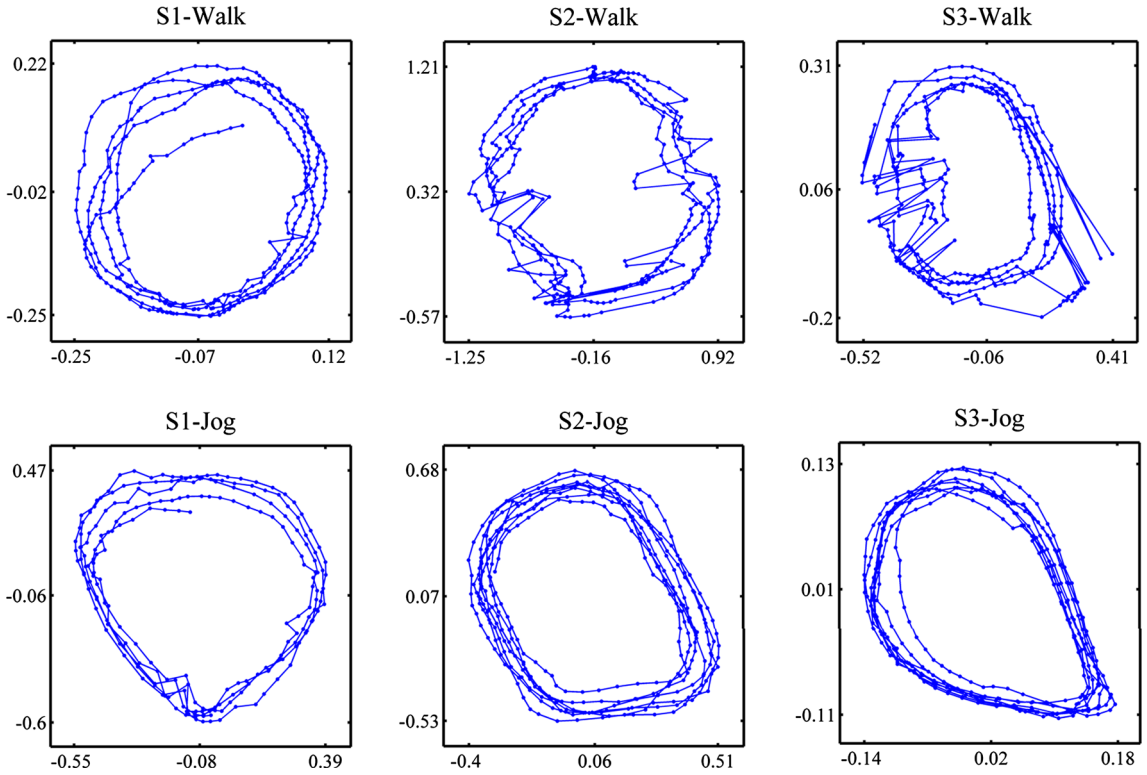


Fig. 5 Low-dimensional pose representations learned using training data for each sequence

6 Empirical study

6.1 Setup

Dataset The performance of the proposed approach is evaluated using real-life benchmark dataset *HumanEva* [20]. The dataset contains multiple subjects performing a set of predefined actions. Originally, for each subject and action the dataset is divided into training, validation and testing sequences. However, all testing sequences are not publicly available. Therefore, very often a different data division is utilized to perform an evaluation, e.g., see [5].

In the experiment we focused on two motion types, namely, *walking* and *jogging*, performed by three different persons, i.e., S1, S2, S3, which results in six various sequences. In each sequence we used 350 and 300 frames from different training trials for training and validation sets, respectively. Only the sequence S1-Jog contained 200 and 200 frames in training and validation sets, respectively. For testing we utilized first 200 frames from each validation trial.

Evaluation methodology The aim of the experiment is to evaluate the proposed approach using MRPF. The presented method was tested against two well-known approaches using the ordinary sampling importance resampling (SIR) and the annealed particle filter (APF). These two methods are usually used as baselines for comparison on *HumanEva*, e.g., [5,13]. The code of these approaches is provided together with *HumanEva*. In both methods gaussian diffusion as the dynamics model was applied.

Each motion sequence is synchronized with measurements from the *MOCAP* system and thus it is possible to evaluate the difference between the true values of a pose configuration with the estimated ones using the following equation ($\mathbf{w}(\cdot) \in \mathcal{W}$ denotes M points on a body for given state variables):

$$\operatorname{err}(\hat{\mathbf{x}}_{1:T}) = \frac{1}{TM} \sum_{t=1}^{T} \sum_{\mathbf{w} \in \mathcal{W}} \|\mathbf{w}(\mathbf{x}_t) - \mathbf{w}(\hat{\mathbf{x}}_t)\|. \tag{31}$$

The obtained value of the error $\operatorname{err}(\hat{\mathbf{x}}_{1:T})$ is expressed in millimeters.

In the empirical study we used the following number of particles: (i) MRPF with 500 particles, (ii) SIR with 500 particles, and (iii) APF with 5 annealing layers with 100 particles each. The low-dimensional manifold had 2 dimensions. In Fig. 5 the learnt low-dimensional embeddings are presented. All parameters (except $\gamma_{\mathbf{x}} = 10^{-4}$) were set according to the optimization process. The methods were run 5 times.

6.2 Results and discussion

The averaged results obtained within the experiment are gathered in Table 1. The results show that the proposed approach

Table 1 The tracking errors $\mathrm{err}(\hat{\mathbf{x}}_{1:T})$ for all methods are expressed as an average and a standard deviation (in brackets)

Sequence	APF	SIR	MRPF
S1-Walk	107 (31)	82 (18)	**69 (7)**
S1-Jog	111 (17)	**81 (4)**	82 (8)
S2-Walk	106 (16)	95 (7)	**86 (12)**
S2-Jog	121 (9)	106 (13)	**94 (8)**
S3-Walk	114 (27)	88 (13)	**79 (10)**
S3-Jog	111 (27)	117 (29)	**70 (8)**

The best results are in bold

with MRPF gave the best results except the sequence S1-Jog for which SIR was slightly better. It is probably caused by the low-quality of this sequence which resulted in shorter training and validation sets. Because of this fact the manifold was possibly not fully discovered.

The worst performance was obtained by the APF. The explanation for such result can be as follows. First, the likelihood model used in the experiment is highly noised by the low quality of the silhouettes achieved in the background subtraction process. The noise in the likelihood model leads to displacement of extrema and thus wrong tracking. Second, the number of particles can be insufficient. However, consid-

Table 2 The tracking errors $\mathrm{err}(\hat{\mathbf{x}}_{1:T})$ for different body parts are expressed as an average and a standard deviation (in brackets)

Seq.	Algorithm	Body part			
		Torso	Head	Legs	Arms
S1-Walk	APF	50 (5)	46 (5)	110 (34)	133 (63)
	SIR	41 (4)	**39 (2)**	74 (28)	111 (32)
	MPF	**36 (1)**	40 (4)	**69 (13)**	**85 (7)**
S1-Jog	APF	44 (6)	49 (7)	96 (21)	158 (30)
	SIR	37 (3)	**40 (3)**	**63 (8)**	122 (17)
	MPF	**35 (4)**	43 (6)	80 (19)	**106 (7)**
S2-Walk	APF	73 (6)	78 (5)	120 (35)	107 (11)
	SIR	76 (5)	76 (3)	101 (9)	98 (11)
	MPF	**61 (6)**	**72 (4)**	**97 (10)**	**85 (20)**
S2-Jog	APF	66 (9)	77 (3)	106 (12)	161 (15)
	SIR	62 (7)	**73 (2)**	**86 (8)**	146 (40)
	MPF	**57 (2)**	75 (4)	102 (9)	**101 (11)**
S3-Walk	APF	48 (3)	43 (7)	143 (35)	120 (42)
	SIR	53 (10)	44 (10)	**93 (17)**	104 (14)
	MPF	**39 (4)**	**35 (3)**	99 (26)	**81 (11)**
S3-Jog	APF	50 (13)	54 (14)	105 (15)	148 (51)
	SIR	45 (11)	51 (11)	96 (22)	172 (48)
	MPF	**38 (3)**	**40 (3)**	**74 (4)**	**81 (15)**

The best results are in bold

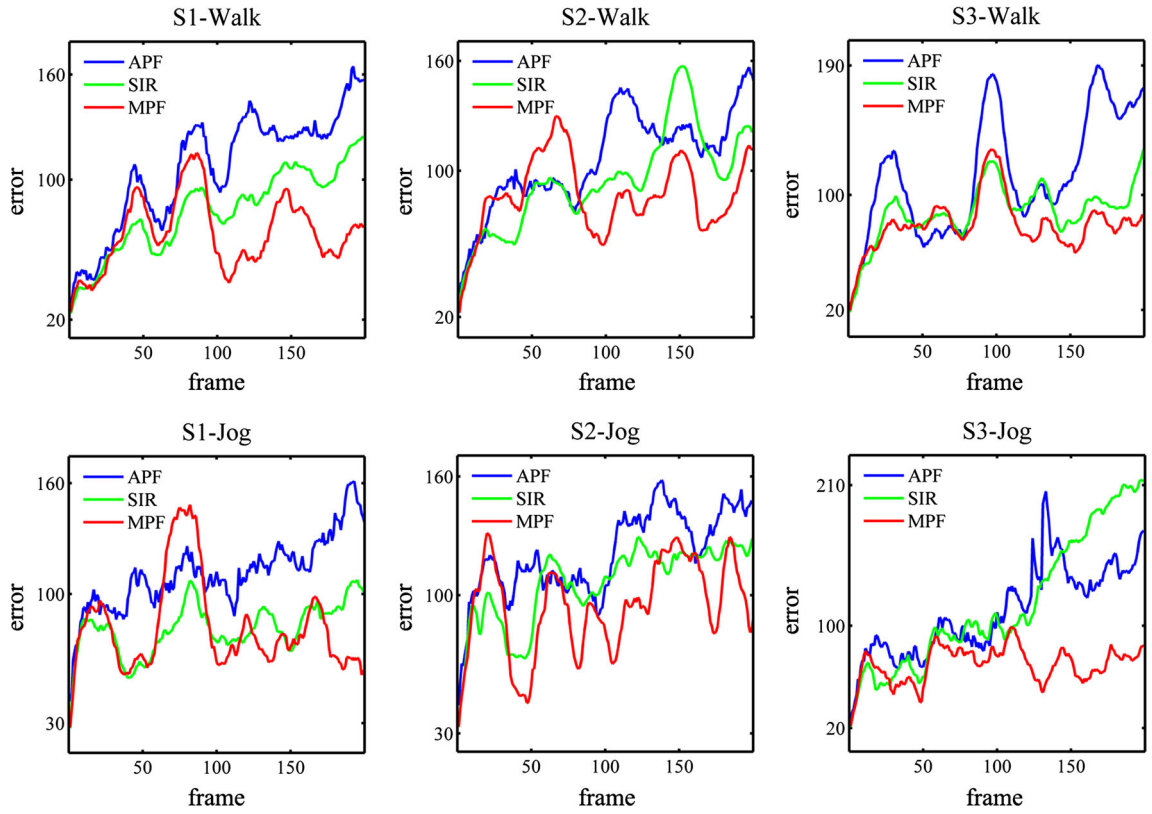
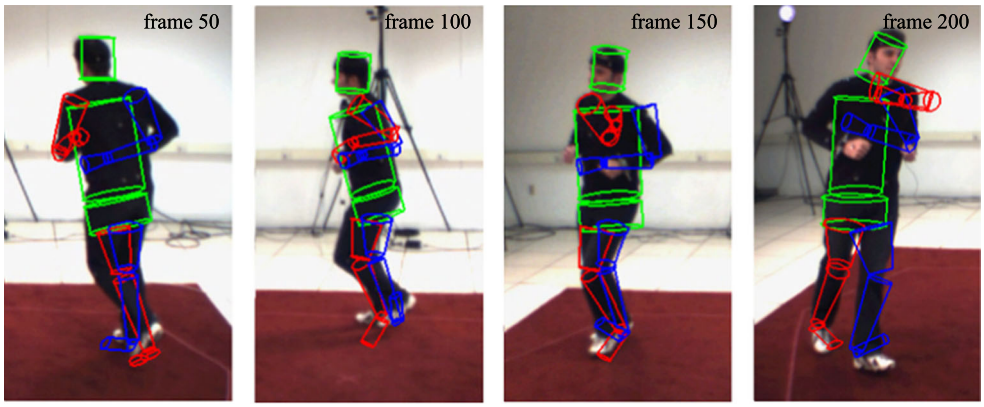
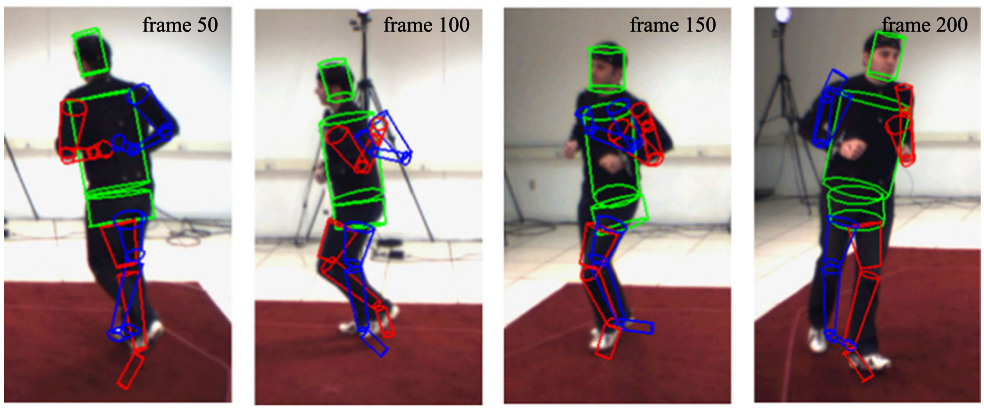


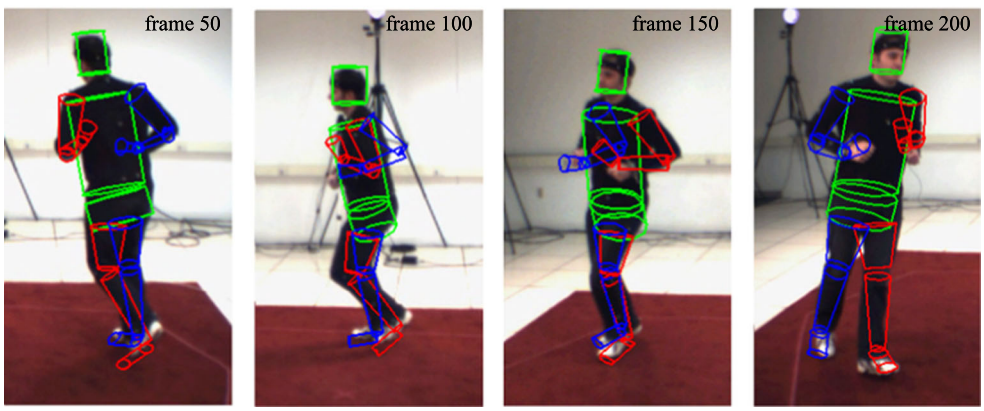
Fig. 6 Tracking error rate over the test sequences



(a) Particle Filter (SIR)



(b) Annealed Particle Filter (APF)



(c) Manifold Regularized Particle Filter (MPF)

Fig. 7 Example frames from S3-Jog test sequence

ering larger number of particles would cause prohibitively long execution time.

In Fig. 6, the tracking error rates are presented. We can see there that the MPRF method behaves more stable than SIR and APF, i.e., the error is not accumulating over consecutive frames. This is the effect of keeping the trajectory close to the manifold. It is especially important if the image evidence is poor, which leads to huge ambiguity. For more detailed consideration on this problem see Table 2, where individual tracking errors for different body parts are presented. Notice that MRPF always achieves better results for arms, where the ambiguity is the biggest since in most of the tracking time arms stay cluttered by the torso. This effect can be also seen in Fig. 7, where some example frames from the last test sequence are shown.

7 Conclusions

In this paper, a fully Bayesian approach to the articulated human motion tracking was proposed. The modification of the standard Condensation algorithm is based on introducing low-dimensional manifold as a regularizer that incorporates prior knowledge about the specificity of human motion. The application of the low-dimensional manifold allows to restrict the space of possible pose configurations. The idea is based on the application of GPLVM with back constraints. At the end of the paper, the experiment was carried out using the real-life benchmark dataset *HumanEva*. The proposed approach was compared with two particle filters, namely, SIR and APF, and the obtained results showed that it outperformed both of them.

Acknowledgments The research conducted by the authors has been partially co-financed by the Ministry of Science and Higher Education, Republic of Poland, namely, Adam Gonczarek: Grant No. B50137W8/K3, Jakub M. Tomczak: Grant No. B50106W8/K3.

References

1. Agarwal, A., Triggs, B.: Recovering 3d human pose from monocular images. Pattern Anal. Mach. Intell. IEEE Trans. **28**(1), 44–58 (2006)
2. Andriluka, M., Roth, S., Schiele, B.: Monocular 3D pose estimation and tracking by detection. In: CVPR '10 Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (2010)
3. Bo, L., Sminchisescu, C.: Twin Gaussian processes for structured prediction. Int. J. Comput. Vis. **87**, 28–52 (2010)
4. Daubney, B., Gibson, D., Campbell, N.: Real-time pose estimation of articulated objects using low-level motion. In: CVPR '08 Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (2008)
5. Daubney, B., Xie, X.: Tracking 3D human pose with large root node uncertainty. In: CVPR '11 Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (2011)
6. Deutscher, J., Reid, I.: Articulated body motion capture by stochastic search. Int. J. Comput. Vis. **61**(2), 185–205 (2005)
7. Doucet, A., Johansen, A.M.: A tutorial on particle filtering and smoothing: fifteen years later. In: Crisan, D., Rozovsky, B. (eds.) Handbook of Nonlinear Filtering, pp. 656–705. Oxford University Press, London, U.K, (2011)
8. Isard, M., Blake, A.: CONDENSATION-conditional density propagation for visual tracking. Int. J. Comput. Vis. **29**(1), 5–28 (1998)
9. Ji, X., Liu, H.: Advances in view-invariant human motion analysis: a review. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **40**(1), 13–24 (2010)
10. Kanaujia, A., Sminchisescu, C., Metaxas, D.: Semi-supervised hierarchical models for 3d human pose reconstruction. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pp. 1–8 (2007)
11. Lawrence, N.: Probabilistic non-linear principal component analysis with gaussian process latent variable models. J. Mach. Learn. Res. **6**, 1783–1816 (2005)
12. Lawrence, N.D., Quiñonero-Candela, J.: Local distance preservation in the gp-lvm through back constraints. In: Proceedings of the 23rd International Conference on Machine Learning, ACM, pp. 513–520 (2006)
13. Li, R., Tian, T.P., Sclaroff, S., Yang, M.H.: 3D human motion tracking with a coordinated mixture of factor analyzers. Int. J. Comput. Vis. **87**(1–2), 170–190 (2010)
14. Memisevic, R., Sigal, L., Fleet, D.J.: Shared kernel information embedding for discriminative inference. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 2852–2859 (2009)
15. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. **104**(2), 90–126 (2006)
16. Petersen, K.B., Pedersen M.S.: The matrix cookbook. Technicalreport. Technical University of Denmark (2012). http://www2.imm.dtu.dk/pubdb/p.php?3274
17. Peursum, P., Venkatesh, S., West, G.: A study on smoothing for particle-filtered 3D human body tracking. Int. J. Comput. Vis. **87**, 53–74 (2010)
18. Poppe, R.: Vision-based human motion analysis: an overview. Comput. Vis. Image Underst. **108**(1), 4–18 (2007)
19. Rasmussen, C.E., Williams, C.K.: Gaussian Processes for Machine Learning. The MIT Press, Cambridge (2006)
20. Sigal, L., Balan, A.O., Black, M.J.: HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int. J. Comput. Vis. **87**(1–2), 4–27 (2010)
21. Sigal, L., Bhatia, S., Roth, S., Black, M.J., Isard, M.: Tracking loose-limbed people. In: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol. 1, pp. I-421 (2004)
22. Singh, V.K., Nevatia, R., Huang, C.: Efficient inference with multiple heterogeneous part detectors for human pose estimation. In: ECCV'10 Proceedings of the 11th European Conference on Computer Vision Conference on Computer Vision: Part III, pp. 314–327 (2010)
23. Sun, M., Telaprolu, M., Lee, H., Savarese, S.: An efficient branch-and-bound algorithm for optimal human pose estimation. In: CVPR '12 Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (2012)
24. Taylor, G.W., Sigal, L., Fleet, D.J., Hinton, G.E.: Dynamical binary latent variable models for 3D human pose tracking. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 631–638 (2010)
25. Tian, T.P., Li, R., Sclaroff, S.: Tracking human body pose on a learned smooth space. Tech. Rep. 2005–029, Boston University Computer Science Department (2005)
26. Urtasun, R., Fleet, D.J., Fua, P.: 3D people tracking with gaussian process dynamical models. In: CVPR '06 Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition (2006)
27. Urtasun, R., Fleet, D.J., Hertzmann, A., Fua, P.: Priors for people tracking from small training sets. In: ICCV '05 Proceedings of the Tenth IEEE International Conference on Computer Vision, pp. 403–410 (2005)
28. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. Pattern Anal. Mach. Intell. IEEE Trans. **30**(2), 283–298 (2008)
29. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 1385–1392 (2011)

Adam Gonczarek is an Assistant Professor in Department of Computer Science at Wrocław University of Technology. He received a M. Sc. degree in Computer Science from Wrocław University of Technology, Faculty of Computer Science and Management, Poland, in 2009, and a M.Sc. degree in Mathematics from Wrocław University of Technology, Faculty of Fundamental Problems of Technology, Poland, in 2010. In 2013 he obtained a Ph.D. degree in Computer Science (with honours) from Wrocław University of Technology, Poland. His Ph.D. research area was focused on pose estimation and motion tracking using generative models and Bayesian learning. Currently, his research interests include machine learning and computer vision, with special concern on deep learning and biomedical applications.

Jakub M. Tomczak is an Assistant Professor in Department of Computer Science at Wrocław University of Technology. He received a double M.Sc. degree in Computer Science within a Double Diploma Programme from Wrocław University of Technology, Poland, and Blekinge Institute of Technology, Sweden, in 2009. In 2013 he obtained a Ph.D. degree in Computer Science (with honours) from Wrocław University of Technology, Poland. His research interests focus on machine learning, with special concern on deep learning, variational inference and Bayesian learning in application to image analysis and medical domain.